REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 25-03-2008	**2. REPORT TYPE** Final Technical Report	**3. DATES COVERED** *(From - To)* 12/13/2004 - 12/31/2007

4. TITLE AND SUBTITLE Optimal Mapping when Datasets are Massive	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER N00014-05-1-0133
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Cressie, Noel, A.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The Ohio State University Research Foundation 1960 Kenny Road Columbus, OH 43210-1063	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research 875 North Randolph Street Arlington, VA 22203-1995	**10. SPONSOR/MONITOR'S ACRONYM(S)** ONR
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Maps are an extremely important part of any military operation and producing timely and accurate maps is an essential planning tool. Ocean-floor or terrain data are static and spatial, while meteorological or visible data are dynamic and spatio-temporal. Data upon which maps are based can be simultaneously massive and sparse, and they are noisy. In the presence of uncertainty due to missingness and measurement-error noise, spatial and spatio-temporal statistical analysis of massive datasets is challenging. The massiveness causes problems in computing optimal spatial predictors, such as kriging, since one has to solve (and store) systems of equations equal to the size of the data. In addition, a large spatial domain is often associated with non-stationary behavior over that domain. These problems are solved using statistical methodology, developed under the grant, called Fixed Rank Kriging.

15. SUBJECT TERMS

Basis functions; Bayesian analysis; kriging; spatial statistics; spatio-temporal statistics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*

FINAL TECHNICAL REPORT
"OPTIMAL MAPPING WHEN DATASETS ARE MASSIVE"

Grant No.:	N00014-05-1-0133
Period of Award:	12/31/2004-12/31/2007
PI Name:	Noel Cressie
Address:	Department of Statistics The Ohio State University 1958 Neil Avenue Columbus OH 43210-1247
Phone Number:	614-292-5194
Fax Number:	614-292-2096
e-mail Address:	ncressie@stat.osu.edu

"OPTIMAL MAPPING WHEN DATASETS ARE MASSIVE"

Objectives

Maps are an extremely important part of any military operation and producing timely and accurate maps is an essential planning tool. Ocean-floor or terrain data are static and spatial, while meteorological or visibility data are dynamic and spatio-temporal. Data upon which maps are based can be simultaneously massive and sparse, and they are noisy. In the presence of uncertainty due to missingness and measurement-error noise, spatial and spatio-temporal statistical analysis of massive datasets is challenging. The massiveness causes problems in computing optimal spatial predictors, such as kriging, since one has to solve (and store) systems of equations equal to the size of the data. In addition, a large spatial domain is often associated with non-stationary behavior over that domain. The objectives are: (1) construct a flexible family of non-stationary covariance functions using a truncated set of basis functions, fixed in number; (2) develop the necessary methodology and algorithms for covariance-parameter estimation; (3) derive optimal spatial or spatio-temporal maps that account for uncertainties statistically; and (4) incorporate spatial and spatio-temporal dependencies into the analysis of sensor-network data.

Impact/Applications

The US Navy has great need for statistical processing to produce current maps and to forecast spatial fields in a rapidly changing environment. The massive-dataset-mapping technology called Fixed Rank Kriging (FRK), has now been published along with several applications. The main paper for spatial prediction, by Cressie and Johannesson, has been published in 2008 in the *Journal of the Royal Statistical Society*, one of the top three statistics journals in the world. A paper on spatio-temporal prediction applied to remote-sensing data from satellites, by Shi and Cressie, has been published in 2007 in *Environmetrics*. The statistical models of dependence are highly flexible and the computational algorithm is extremely fast, providing the mapping community with the means of making complete maps of both the surface (kriging) and its uncertainty (kriging variance).

We have developed two web sites at The Ohio State University:
www.stat.osu.edu/~C2 considers probability and statistics in Command and Control.
www.stat.osu.edu/~sses/research_mds.html shows the FRK mapping approach that fills in missing data and smoothes out noise. The application given there is to global mapping of aerosol optical depth data obtained from the MISR instrument on the Terra satellite.

During the period of the grant, a two-year contract was signed with Oak Ridge National Laboratory (ORNL) to incorporate spatial and spatio-temporal dependencies into the analysis of sensor-network data. This allowed a postdoctoral fellow to be supported jointly by this ONR grant and by the ORNL contract

"OPTIMAL MAPPING WHEN DATASETS ARE MASSIVE"

Personnel

Principal Investigator:	Noel Cressie, PhD
Postdoctoral Fellow:	Chungfeng Huang, PhD (partial ONR support)
Research Assistants:	Yonggang Yao Hongfei Li (partial ONR support) Lei Kang

Technical Approach

This research is focused on mapping from data observed at many locations distributed through space. The spatial dependence is captured through a set of r (not necessarily orthogonal) basis functions,

$$\mathbf{S}(\mathbf{u}) \equiv (S_1(\mathbf{u}),\ldots, S_r(\mathbf{u}))' ; \quad \mathbf{u} \in \mathbb{R}^d , \tag{1}$$

where r is fixed and the setting is a d-dimensional Euclidean space. For any $r\times r$ positive-definite matrix $\mathbf{K}$, the covariance function between two elements $Y(\mathbf{u})$ and $Y(\mathbf{v})$ of the spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ is assumed to be,

$$C(\mathbf{u},\mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}) . \tag{2}$$

Associated with the covariance function (2) is a spectral representation for $Y(\cdot)$, given by:

$$Y(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\upsilon} ; \quad \mathbf{s} \in D , \tag{3}$$

where $\boldsymbol{\upsilon}$ is an r-dimensional random vector such that $\mathrm{var}(\boldsymbol{\upsilon}) = \mathbf{K}$. We assume that we have n measurements of $Y(\cdot)$ that provide data $\mathbf{Z} \equiv (Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n))'$ at locations $\mathbf{s}_1,\ldots,\mathbf{s}_n$, where

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i) ; \quad i = 1,\ldots,n , \tag{4}$$

and $\varepsilon(\cdot)$ is a mean zero, variance σ_ε^2, white-noise measurement process independent of $Y(\cdot)$.

By using the flexible model (2) and Fixed Rank Kriging, which is an optimal-spatial-prediction methodology (Cressie and Johannesson, 2008), massive datasets can be processed in $O(n)$ flops, and inversion of only an $r\times r$ matrix is needed for each spatial-prediction location; recall that r is fixed.

FRK relies fundamentally on inversion of the $n \times n$ covariance matrix of the data, namely

$$\mathrm{var}(\mathbf{Z}) = \mathbf{SKS}'+\sigma_\varepsilon^2\mathbf{I},$$

where $\mathbf{S}$ is the $n\times r$ matrix $(\mathbf{S}(\mathbf{s}_1)\ \dots\ \mathbf{S}(\mathbf{s}_n))'$ and $\mathbf{I}$ is the $n\times n$ identity matrix. Under normal circumstances, inversion of an $n\times n$ covariance matrix takes $O(n^3)$ flops. FRK reduces this to $O(n)$ by invoking the following formula:

$$(\mathrm{var}(\mathbf{Z}))^{-1} = (1/\sigma_\varepsilon^2)\mathbf{I} - (1/\sigma_\varepsilon^4)\mathbf{S}\{\mathbf{K}^{-1} + (1/\sigma_\varepsilon^2)\mathbf{S}'\mathbf{S}\}^{-1}\mathbf{S}',$$

where all matrix inverses are of $r\times r$ matrices and r is fixed.

The temporal component can be incorporated as follows. Let $\{Y(\mathbf{s};t) : \mathbf{s} \in D \subset ☼^d, t = 1,2,\dots\}$ denote a process to be sensed; suppose that data are obtained at locations $\mathbf{s}_1,\dots,\mathbf{s}_n$, and at successive times $t = 1,2,\dots$. Using the same terminology as for the spatial case, we wish to predict $Y(\cdot;t_0)$ based on data,

$$\mathbf{Z}(t) \equiv (Z(\mathbf{s}_1;t),\dots,Z(\mathbf{s}_n;t))'; \quad t = 1,\dots,t_0\,. \tag{5}$$

Adding the time component also allows us to pose the problem of *forecasting* $Y(\cdot;t_0 + 1)$. Key to obtaining solutions to the prediction and forecasting problems is the assumption of a spatio-temporal model.

We generalize (3) by assuming that

$$Y(\mathbf{s};t) = \mathbf{S}(\mathbf{s})'\boldsymbol{\upsilon}(t)\,; \quad \mathbf{s} \in D, \quad t = 1,2,\dots, \tag{6}$$

where $\{\boldsymbol{\upsilon}(t) : t = 1,2,\dots\}$ is a temporal stochastic process. For example, if $\boldsymbol{\upsilon}(1),\boldsymbol{\upsilon}(2),\dots$ are independent and identically distributed, the data $\mathbf{Z}(1),\mathbf{Z}(2),\dots,\mathbf{Z}(t_0)$ could be viewed as independent realizations of the spatial function $Y(\cdot)$ given by (3). Another example that we find very interesting is that of early detection of bioterrorism events, where we assume a dynamic model for $\boldsymbol{\upsilon}(\cdot)$. Wikle and Cressie (1999) did this in a climate context, and their basis functions were empirical orthogonal functions. The model (6) is more general and offers the oportunity of early detection by continually testing for a regime shift in the dynamic model for $\boldsymbol{\upsilon}(\cdot)$.

References

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **70**, 1-18.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815-829.

"OPTIMAL MAPPING WHEN DATASETS ARE MASSIVE"

Results

Listed below are the presentations and publications related to the grant.

Presentations

2005

Invited seminar speaker, Department of Statistics, North Carolina State University, Raleigh, NC; "A fast, optimal spatial prediction method for massive datasets", January 2005.

Invited seminar speaker, Mathematics Laboratory, Universite de Paris Sud, Orsay, France; "A fast, optimal spatial-prediction method for massive datasets", March 2005.

Invited seminar speaker, Division of Mathematical and Information Sciences, CSIRO, Perth, Australia; "Geostatistical prediction of spatial extremes and their extent", April 2005.

Invited seminar speaker, Department of Statistics, University of Padova, Padova, Italy; "Dynamic multi-resolution spatial models", June 2005.

Presented an invited paper at SAMSI Workshop on Bridging Statistical Approaches and Sequential Data Assimiliation, Research Triangle Park, NC; "Data assimilation using multi-resolution spatio-temporal models", June 2005.

Co-authored a contributed poster (with J. Zhang and P. Craigmile) at Joint Statistical Meetings, Minneapolis, MN; "Predicting exceedance regions for geostatistical processes", August, 2005.

2006

Presented a keynote address to 2006 White Conference on Mastering the Data Explosion in the Earth and Environmental Sciences, Australian Academy of Sciences, Canberra, Australia; "Spatial prediction for massive datasets", April 2006.

Dan and Carol Burack President's Distinguished Lecturer, University of Vermont, Burlington, VT. Presented Burack lecture; "Massive but sparse spatial data", May 2006.

Presented an invited paper (with T. Shi) at Second NASA Data Mining Workshop: Issues and Applications in Earth Sciences, Pasadena, CA; "Satellite data: Massive but sparse", May 2006.

Presentations, ctd.

Presented an invited paper (with G. Johannesson) at American Statistical Association Annual Meeting, Seattle, WA; "Fixed rank kriging for massive datasets". Also co-authored a contributed paper (with C. Huang, Y. Yao, and T. Hsing); "Cokriging with generalized cross-covariances for detecting radioactivity". Also coauthored a contributed paper (with Y. Yao); "Spatial multivariate EOFs: Discrete to continuous approximations". Also co-authored a contributed paper (with J. Zhang and P. Craigmile); "Predicting spatial exceedance regions". Also co-authored a contributed paper (with H. Li and C. Calder); "Testing for spatial dependence based on the SAR model", August 2006.

Presented the keynote address at METMA3, International Workshop on Spatio-Temporal Modelling, Pamplona, Spain; "Spatio-temporal satellite data processing", September 2006.

Presented an invited paper at the International Symposium on Statistical Analysis of Spatio-Temporal Data, Tokyo, Japan; "Spatio-temporal satellite data processing", November 2006.

2007

Invited seminar speaker, Institute of Statistics and Decision Sciences, Duke University, Durham, NC; "Optimal spatial prediction for large spatial datasets", February 2007.

Invited seminar speaker, SAMOS, Universite de Paris 1 (Sorbonne), France; "Spatial prediction for massive datasets", March 2007.

Presented five lectures as Principal Lecturer, 32nd Spring Lecture Series on Spatial and Spatio-Temporal Statistics, University of Arkansas, Fayetteville, AR, April 2007.

Presented an invited paper at Workshop Spatial Statistics, Universite de Paris 1 (Sorbonne), France; "Predicting spatial exceedance regions", April 2007.

Co-authored a contributed paper (with C. Huang and T. Hsing at American Statistical Association Annual Meeting, Salt Lake City, UT); "Spectrum estimation for isotropic intrinsically stationary spatial processes". Also co-authored a contributed paper (with H. Li and C. Calder); "Exploratory spatial data analysis using APLE statistics". Also co-authored a contributed paper (with L. Kang and D. Liu); "Spatial statistical analysis of doctors' prescription amounts by region". Also co-authored a contributed paper (with T. Shi); "Spatio-temporal processing of MISR's aerosol optical depth data", July 2007.

Presented an invited paper (with A. Braverman and H. Nguyen) at 2007 American Geophysical Union Fall Meeting, San Francisco, CA; "Fusing measurements statistically: Combining aerosol data from MISR and MODIS", December 2007.

"OPTIMAL MAPPING WHEN DATASETS ARE MASSIVE"

Publications: Refereed Articles

Craigmile, P. F., Cressie, N., Santner, T. J., and Rao, Y. (2006). Bayesian inference on environmental exceedances and their spatial locations. *Extremes*, **8**, 143-159.

Cressie, N. (2006). Block kriging for lognormal spatial processes. *Mathematical Geology*, **38**, 413-443.

Cressie, N. and Verzelen, N. (2007). Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields. *Computational Statistics and Data Analysis*, **52**, 2794-2807.

Li, H., Calder, C. A., and Cressie, N. (2007). Beyond Moran's I: Testing for spatial dependence based on the SAR model. *Geographical Analysis*, **39**, 357-375.

Sain, S. and Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*,**140**, 226-259.

Shi, T. and Cressie, N. (2007). Global statistical analysis of MISR aerosol data: A massive data product from NASA's Terra satellite. *Environmetrics*, **19**, 665-680.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial datasets. *Journal of the Royal Statistical Society, Series B*, **70**, 1-18.

Cressie, N. and Kapat, P. (2008). Some diagnostics for Markov random fields. *Journal of Computational and Graphical Statistics*, forthcoming.

Zhang, J., Craigmile, P.F., and Cressie, N. (2008). Loss function approaches to predict a spatial quantile and its exceedance region. *Technometrics*, forthcoming.

Publications: Non-refereed articles

Cressie, N. and Yao, Y. (2005). Release of *Web-Project: TCO*, showing spatial prediction of total column ozone over the globe using a fast multi-resolution spatial statistical model (http://ww.stat.osu.edu/~sses/collab_ozone.php).

Ganguly, A.R., Hsing, T., Katz, R., Erickson, D.J., Ostrouchov, G., Wilbanks, T.J., and Cressie, N. (2005). Multivariate dependence among extremes, abrupt change, and anomalies in space and time for climate applications, in *Proceedings of the International Workshop on Data Mining Methods for Anomaly Detection*, eds D. Margineantu, S. Bay, P. Chan, and T. Lane, 25-26.

Publications: Non-refereed articles, ctd.

Cressie, N. and Johannesson, G. (2006). Spatial prediction for massive datasets, in *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*. Australian Academy of Science, Canberra, Australia (11 pp.)

Calder, C. and Cressie, N. (2007). Some topics in convolution-based spatial modeling, in *Proceedings of the 56th Session of the International Statistical Institute*, Lisbon, Portugal, forthcoming.

Sain, S.R., Furrer, R., and Cressie, N. (2007). Combining regional climate model output via a multivariate Markov random field model, in *Proceedings of the 56th Session of the International Statistical Institute*, Lisbon, Portugal, forthcoming.

Articles submitted/in preparation

Huang, C., Cressie, N., Yao, Y., and Hsing, T. (2007). Multivariate intrinsic random functions for cokriging, under revision for *Mathematical Geosciences*.

Huang, C., Hsing, T., and Cressie, N. (2007). On a general-spline estimator for the spectral density function, under journal review.

Huang, C., Hsing, T., Cressie, N., Ganguly, A.R., Protopopescu, V.A., and Rao, N.S. (2007). Statistical analysis of plume model identification based on sensor network measurements, under revision for *Transactions on Sensor Networks.*

Kang, L., Liu, D., and Cressie, N. (2007). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models, under journal review.